Shallow-transfer rule-based machine translation from Czech to Polish

Joanna Ruth[1] Jimmy O'Regan[2]

[1]Gdańsk University of Technology
joannaruth1@gmail.com

[2]Eolaistriu Technologies
joregan@gmail.com

Why Czech?

- Budweiser
- Pilsner

Why Czech?

- Budweiser → Budějovice.
- Pilsner → Plzeň.

Czech is the language of beer!

Some Famous Czechs

- Dvořák
- Jan Hus
- Kafka
- Good King Wenceslas
- Eva Herzigová
- Petra Nemčová
- Ivana Trump

Some Famous Poles

- Chopin
- Pope John Paul II
- Copernicus
- Marie Curie (Maria Skłodowska)
- Ludwik Zamenhof (Esperanto)
- Joseph Conrad
- Roman Polański

Czech and Polish

> *In the 10th century, Czech and Polish were still basically the same language, which then began to diverge from each other, but even until the 14 century, Czechs and Poles understood each other without problems.*

Czech Wikipedia, *Polština*

Czech and Polish

Both Western Slavic languages:
Czech:

- 12 million speakers
- Czech word in English: robot

Polish:

- 50 million speakers
- Polish word in English: vodka

Czech and Polish: Similarities

- Medium inflected:

Czech and Polish: Similarities

- Medium inflected:
  - 7 cases
  - 3 genders
  - Animacy distinction

Czech and Polish: Similarities

- Medium inflected:
  - 7 cases
  - 3 genders
  - Animacy distinction
- Relatively free word order:

Czech and Polish: Similarities

- Medium inflected:
  - 7 cases
  - 3 genders
  - Animacy distinction
- Relatively free word order:
  - Ala ma kota
  - Kota Ala ma
  - Ala kota ma
  - Ma Ala kota
  - . . .

Czech and Polish: Cases

Case	Czech	Polish
Nominative	matka	matka
Genitive	matky	matki
Dative	matce	matce
Acusative	matku	matkę
Instrumental	matkou	matką
Locative	matce	matce
Vocative	matko	matko

Czech and Polish: NP Differences

	Czech	Polish
Word order	adj before noun	adj before or after noun
Possessive	adjectival form	genitive

Czech and Polish: VP Differences

	Czech	Polish
"ought to"	by + mít_{past}+*INF*	powinien + INF
"while *x*-ing"	present transgressive (adj)	adverb *(-jąc)*
"having *x*-ed"	past transgressive (adj)	adverb *(-wszy)*
past tense	personal form from *být*	conjugated

Lexical differences: A little history

(“Accidental”) Germanisation of Bohemia began in 1620. Czech ceased to exist as a literary language.
Poland was partitioned in the 18th century. Germanisation began in the Prussian partition.
However:
Publication allowed in the Austro-Hungarian and Russian partitions, and in France. Polish continued to thrive as a literary language.

Lexical differences: Czech Revival

Czech was revived in the 18th and 19th Centuries.
Jungmann's dictionary was partly based around the Bible of Kralice (16th Century), with German words replaced by Slavic (Russian, Bulgarian) loans and neologisms.
This lead to an increase in the lexical differences between Czech and Polish.

Czech vs. Polish: Viewpoints

The Czechs and Poles are neighbours, and have less-than-flattering views of each other.

Polish view of Czech: Child-like

- More lexicalised diminutives.
- Loss of palatalisation.

i.e., spoken Czech sounds a little like Polish babytalk

Czech view of Polish: Archaic

- Digraphs (sz, cz) instead of caron.
- Retention of Proto-Slavic “nasal vowels”.

i.e., written Polish looks a little like early written Czech.

Czech View of Polish

"In Poland, a comical lisping language is spoken, dominated by different variants of the sound 'sh'. Polish has 17 species of them and the exact pronunciations are not known by the Poles themselves. . . . The current pronunciation of the Polish language only stabilised during World War II. . . . To avoid German attacks, it could not be distinguished from static."

"V Polsku se mluví komickým šišlavým jazykem, ve kterým prevládají ruzný varianty hlásky 'š'. Polština jich má 17 druhu a jejich presnou výslovnost neznají ani sami Poláci. . . . Soucasná výslovnost polského jazyka se ustálila teprve až behem 2. svetové války. . . . Aby nebylo pred Nemci nápadné, nesmelo být odlišitelné od statického šumu."

http://necyklopedie.wikia.com/wiki/Polsztyna

An aside: “l-participle”

The Czech past form is sometimes referred to as the “l-participle”. Whether or not it’s a participle is arguable.

- Not fully periphrastic: past.p3 uses no auxiliary.[1]
- Not fully adjectival.
- Not a modifier.

[1] The Sorbian languages do

An aside: The Traditional View of Czech

In addition to the “l-participle”, there are a few other ways in which it was more helpful to avoid the Czech linguistic tradition:

Verbal nouns/adjectives

Typically considered to be entirely lexicalised.

We chose to add them, in anticipation of the Polish → Czech direction; we don’t consider the Czech case any more compelling than verbal substantives in other languages, and we want the data to be useful for future potential language pairs.

Synthetic adjectives

All regular adjectives are considered synthetic.

In reality, analytic constructs using *více/nejvíce* are used with many adjectives to form the comparative/superlative.

Nejexotermičtější → “Exothermicest”

An aside: The Traditional View of Polish

Historically, Polish verbs added an enclitic form of być to the past tense of verbs to express person. This view is still used in Polish linguistics[2]; however, this viewpoint is not widely known (nor, typically, even understood) outside of linguistics.

For that reason, we treat Polish verbs as having a full conjugation in the past, and as having a conditional tense.

For other cases of *być* attachment, we found only the *by* 'family' of conjunctions to be productive in modern, professionally written text, and that segmentation ambiguities possible through this attachment (*goście zabili, kogoś widziała*) sufficiently rare to ignore.

[2] See, for example, **Radziszewski and Śniatowski** Maca – a configurable tool to integrate Polish morphological data, *These proceedings*

Some false friends

Polish	Czech	English
kwiecień	duben	April
szukać	hledat	to look for
Czech	Polish	English
květen	maj	May
šukat	…	…

Czech View of Polish, Reprise

Polacy “šukají” cokolwiek, i to dlaczego jest cztery razy więcej Polaków niż Czechów.[3]

Poláci "šukají" kdeco a výsledkem tak je, že jich je 4x tolik co Čechů.

[3] Przepraszam za mój marny polski.

Why not SMT?

Reviewer's comment:

> *In section 3.4, the reader is told about the existence of a parallel corpus including Czech and Polish. This should be mentioned in the introduction along with the motivation of developing this rule-based system (as opposed to a statistical one).*

Why not SMT?

First and foremost:
This project was funded under Google Summer of Code: it had to produce a piece of Open Source *Software*. Apertium's rules include a programmatic element; SMT would be almost impossible to justify.

Why not SMT?

First and foremost:
This project was funded under Google Summer of Code: it had to produce a piece of Open Source *Software*. Apertium's rules include a programmatic element; SMT would be almost impossible to justify.
Secondly:
It's an Apertium project. 'Nuff said.

Why not SMT? Translation Drift

Original

[He] was seated at the breakfast table.

Polish

Jadł śniadanie.

Polish (translation)

He ate breakfast.

Czech

[S]eděl právě u stolu, na němž se snídávalo.

Czech (translation)

He sat right at the table, at which one breakfasts.

(The verb phrases "jadł śniadanie" and "se snídávalo" *almost* align with each other).

Why not SMT?

Data sparseness
: Compounded by relatively large amount of morphological forms.

Lack of truly parallel text
: Most parallel text are mutual translations.

Lack of true corpora
: JRC Acquis is not a corpus.

Aside: JRC Acquis is not a corpus

That might be considered a “bold statement”.
(It’s not. Ask a corpus linguist.)

To be clear, we’re referring to the *Corpus*, not the DGT’s Translation Memory distribution, which, going by the Moses mailing list, is more often used.

Aside: JRC Acquis is not a corpus

JRC Acquis is:

- a dump of raw text
- full of encoding errors[4]
- not reliably sentence aligned: industry practice is to realign
- not annotated
- not maintained

(For contrast, EuroParl is actively maintained, contains document origin annotation, speaker's original language annotation, and PoS annotation. Unfortunately, it contains neither Czech nor Polish)

[4] At least, for Polish and Czech. YMMV.

Why not SMT?

Hierarchical/Syntax Augmented

No available tree parsers.

Factored models

Seems promising

"Phrase" Based

The only real option

Why not SMT: Factored models

> *One example to illustrate the short-comings of the traditional surface word approach in statistical machine translation is the poor handling of morphology. Each word form is treated as a token in itself.*
>
> *. . .*
>
> *While this problem does not show up as strongly in English – due to the very limited morphological inflection in English – it does constitute a significant problem for morphologically rich languages such as Arabic, German, Czech, etc.*

Factored Translation Models, *Philipp Koehn and Hieu Hoang*, EMNLP 2007, `http://homepages.inf.ed.ac.uk/pkoehn/publications/emnlp2007-factored.pdf`.

Why not SMT: Factored models

The Official Version:

> *We reported on experiments that showed gains over standard phrase-based models, both in terms of automatic scores (gains of up to 2% BLEU), as well as a measure of grammatical coherence. These experiments demonstrate that within the framework of factored translation models additional information can be successfully exploited to overcome some short-comings of the currently dominant phrase-based statistical approach.*

Ibid.

Why not SMT: Factored models

Gains of up to 2% BLEU!

Why not SMT: Factored models

The Unofficial Version:

> *"Factored models don't work."*

– `$WELL_KNOWN_SMT_GUY`

Why not SMT: Factored models

The Unofficial Version:

> *“Factored models don’t work.”*

– `$WELL_KNOWN_SMT_GUY`

Unfortunately, BibTeX does not allow:

```
@PubConversation{wmt2010,
 author     = "$WELL_KNOWN_SMT_GUY",
 topic      = "Factored Models",
 year       =  2010,
 pub        = "Lanigan's Plough"
 where      = "Dublin",
}
```

Why not SMT: “Phrase”-Based

We found nothing in the literature about PBMT between Slavic languages.

Of closest relevance was work on Czech to English SMT, which typically uses “stems” (the first 4 characters of a word) of Czech words, to overcome the problem of morphological complexity.

Though this reduces data sparseness, it loses case information, as well as person information in verbs.

This may be adequate for English, but Polish requires that information.

An aside: SMT Terms

What's the deal with SMT people trying to redefine existing linguistic terms?

Phrase
: a group of words functioning as a single unit in the syntax of a sentence

"phrase"
: whole sequences of words, where the lengths may differ

Stem
: the part of the word that is common to all its inflected variants

"stem"
: the first *n* characters of a word

Ok, stem is fair game

Why not SMT: Google Translate

Google Translate is the only online Czech to Polish PBMT system we found
It is also the only online MT system we found of any kind.

Why not SMT: Google Translate

Although few details are available generally, and none for particular language pairs, we can observe that Google translate:

Strips diacritics

: The presence or absence makes no difference to the translation for Polish and Czech (compare with Spanish: missing diacritics provide a different translation)

Uses English as a pivot

: *Jsem* in Czech translates as the English *I*, instead of Polish *Jestem*

Uses "Stemming"

: Manually stemmed sentences translate the same as unstemmed.

Why not SMT: Google Translate

Uses unchecked online wordlists

Google Dictionary for English-Polish and English-Czech contains many entries from online wordlists, known generally to be poor quality[5][6]. Czech to Polish in Google Translate produces several curious translations that could be explained by the triangulation of these lists.

Google Translate's use of poor quality wordlists in particular convinced us of the importance of not trusting our sources.

[5] Ectaco, slovnik.zcu.cz, slovnyk.org, etc.

[6] I'm a language nerd, I take note of the source when I see bad translations

Why not SMT: Google Translate

On the bright side:
We can infer that Google Translate's Czech to Polish pair was not built using data that's not generally available.

Why not SMT: Google Translate

On the bright side:
We can infer that Google Translate's Czech to Polish pair was not built using data that's not generally available. (That, or it has been buried under the bad publicly available data).

Yet Another Aside

The more astute of you will have noticed that I'm trying to talk about everything vaguely related to the system, and not about the system itself.

Yet Another Aside

The more astute of you will have noticed that I'm trying to talk about everything vaguely related to the system, and not about the system itself.
(In fact, I'll be very surprised if I haven't been heckled for it.)

Yet Another Aside

The more astute of you will have noticed that I'm trying to talk about everything vaguely related to the system, and not about the system itself.
(In fact, I'll be very surprised if I haven't been heckled for it.)
That's because it sucks.

Resources

There are many resources available for Czech and Polish.

Resources

There are many resources available for Czech and Polish.
For a given value of “available”.

Resources: Morphological analysis

Resources available in Apertium:
Parts of a Polish morphological analyser from an unsuccessful attempt at English to Polish.
A corpus-derived Czech morphological analyser, from partial work on Czech to Slovenian.
Other resources:
Morfologik (Polish analysis).
F-Morph (non-free: Czech)
LanguageTool's Czech morphological dictionary (corpus derived).
`ispell` dictionaries.

Resources: ispell

Slavic ispell dictionaries are typically designed according to linguistic principles. (This is also true of, e.g., Baltic dictionaries). This gives us an (almost) 1:1 correspondence between suffix + flags in ispell and Apertium paradigms:

Table: Sample mappings between Polish ispell entries and paradigms.

ispell	Apertium
miłość/MN	miłoś/ć__n
matka/mMN	mat/ka__n
droga/mMN	fla/ga__n

Resources: Bilingual

We collected our own bilingual text for testing purposes. (The open content portion of this collection has been donated to the Open Content Text Corpus project).

Resources: Other

We also had a set of rules for Slovakian to Polish and Polish to Slovakian.
As Slovakian and Czech have almost identical syntax, these rules only needed slight modifications to apply to Czech, though the ruleset was far from complete.

Morphological Analysis and Generation

Most of the sources we had were suitable for analysis, but not for generation.

In addition, in the English-Polish project, we faced the problem of the large number of paradigms required for Polish. This made it difficult to edit, and to fix errors.

At the beginning of this project, we decided to abstract the common portions of the paradigms to both reduce the size of the files, and to make it easier to keep consistency in the event of modification.

Paradigms: Before

```
<pardef n="mat/ka__n">
  <e><p><l>ka</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="nom"/></r></p></e>
  <e><p><l>ki</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="gen"/></r></p></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>kę</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="acc"/></r></p></e>
  <e><p><l>ką</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="ins"/></r></p></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
  <e><p><l>ko</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="voc"/></r></p></e>
</pardef>
<pardef n="dro/ga__n">
  <e><p><l>ga</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="nom"/></r></p></e>
  <e><p><l>gi</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="gen"/></r></p></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>gę</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="acc"/></r></p></e>
  <e><p><l>gą</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="ins"/></r></p></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
  <e><p><l>go</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="voc"/></r></p></e>
</pardef>
```

Figure: Before: two paradigms with slight differences

Paradigms: Before

```
<pardef n="BASE__matka">
  <e><p><l>a</l><r><s n="f"/><s n="sg"/><s n="nom"/></r></p></e>
  <e><p><l>i</l><r><s n="f"/><s n="sg"/><s n="gen"/></r></p></e>
  <e><p><l>ę</l><r><s n="f"/><s n="sg"/><s n="acc"/></r></p></e>
  <e><p><l>ą</l><r><s n="f"/><s n="sg"/><s n="ins"/></r></p></e>
  <e><p><l>o</l><r><s n="f"/><s n="sg"/><s n="voc"/></r></p></e>
</pardef>
<pardef n="mat/ka__n">
  <e><p><l>k</l><r>ka<s n="n"/></r></p><par n="BASE__matka"/></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
</pardef>
<pardef n="dro/ga__n">
  <e><p><l>g</l><r>ga<s n="n"/></r></p><par n="BASE__matka"/></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
</pardef>
```

Figure: Those paradigms redefined in terms of a common base

Bilingual Lexicon: Cognates

We used several methods for cognate induction, including a modified version of the method used in creating the Swedish to Danish translator (we needed to use more than single letter substitutions to account for Czech *č, š, ř* to Polish *cz, sz, rz*).

We found the results to be less promising than in sv-da, which we attribute to the historical divergence in lexicons.

We found that by restraining the cognates by suffix to words of the same origin, or derived through pan-Slavic processes, we greatly increased the accuracy of induced cognates.[7]

[7] Basically the same as **Learning a Translation Lexicon from Monolingual Corpora**, *Philipp Koehn and Kevin Knight*, ACL 2002, Workshop on Unsupervised Lexical Acquisition

Bilingual Lexicon: Wordlists

There are relatively few Czech–Polish wordlists available, and most are relatively low quality.
There are many more wordlists for Czech–English and Polish–English. We were used Scannell's method[8] of cognate induction using lexicons created by triangulation (using `apertium-crossdics`).

[8] **Machine translation for closely related language pairs**, *Kevin P. Scannell*, Proceedings of the Workshop "Strategies for developing machine translation for minority languages" at LREC 2006, Genoa, Italy, May 2006, pp103-107.

Bilingual Lexicon: Wikilinks

We used Fran's Wikipedia interwiki link extraction, with mixed results.
Toponyms were almost perfect; regular nouns were hit and miss (more miss than hit).
We got better results by **not** following redirects, as doing so lead to too many wordpairs that were related, but not translations. (A simple filter based on the presence of animacy on only one side was enough to find problematic masculine nouns, but this is not sufficiently general).
Our intuition is that the size of the target Wikipedia (Polish[9]) is part of the problem: as Wikipedias grow in size, articles tend to "group": single line descriptions of related articles are changed into redirects to a single, more comprehensive article.

[9] At the time of writing, the 5th largest Wikipedia

Bilingual Lexicon: Probabilistic

We trained a statistical machine translation system using Moses (Koehn et al., 2007) on the JRC Acquis Corpus (Ralf et al., 2006), extracting the most probable translations.

Bilingual Lexicon: Probabilistic

> *We trained a statistical machine translation system using Moses (Koehn et al., 2007) on the JRC Acquis Corpus (Ralf et al., 2006), extracting the most probable translations.*

...none of which were directly usable.

Bilingual Lexicon: Validation by Intersection

As we had a large number of wordlists, we wished to get more than cognates from them. Each wordlist taken as a whole was unreliable, but contained many correct translations.

We used the simplistic idea that if a significant number of lists agreed, the translation was adequate.

The wordlists were first filtered to select only candidates whose translations matched in terms of part of speech. We used a threshold of greater than 50 percent: if there were 6 possibilities, 4 had to match to be selected. We also stipulated a minimum set of translation choices of 3, to avoid the possibility that one wordlist contained a subset of another.

Some Numbers

Some Numbers

1 Million and 1

Sixty-six

1 Billion, twenty-five, seventy-five thousand

1 Billion and eight, six, something

Zero

1 Million 1

Twenty-two

Seventy-five

Eleven

Eleven

Ok this is the new order

The New Number Order

Shellac, *New Number Order*

Some Numbers

1 Million and 1

Sixty-six

1 Billion, twenty-five, seventy-five thousand

1 Billion and eight, six, something

Zero

1 Million 1

Twenty-two

Seventy-five

Eleven

Eleven

Ok this is the new order

The New Number Order

Shellac, *New Number Order*

Coverage of the analyser is utterly irrelevant to translation, but I include the numbers to appease a certain dreadlocked morphological analysis fanatic.

Some Numbers: Naive Coverage

Corpus	Running tokens	Known tokens	Coverage
Polish	39,293,427	27,997,757	71.25%
Czech	17,165,777	10,925,926	63.65%

Table: Naïve coverage for both translation directions

Evaluation

Corpus	WER	PWER	Free rides	Unknowns
News Samples	71 %	60 %	5 %	28 %
UDHR	88 %	68 %	0 %	22 %

Table: Evaluation results, apertium-pl-cs.

Evaluation

Corpus	WER	PWER	Free rides	Unknowns
News Samples	71 %	60 %	5 %	28 %
UDHR	88 %	68 %	0 %	22 %

Table: Evaluation results, apertium-pl-cs.

You remember that I said it sucks? 71% is awful.

Evaluation: Google Translate

Corpus	WER	PWER
News Samples	76 %	62 %
UDHR	47 %	32 %

Table: Evaluation results for Google Translate.

Evaluation: Google Translate

Corpus	WER	PWER
News Samples	76 %	62 %
UDHR	47 %	32 %

Table: Evaluation results for Google Translate.

...but at least it's better than Google Translate, for text known not to have been part of Google Translate's training set.

Disambiguation

The main problem is that of disambiguation.

PoS tagging is not a difficult problem, though traditional Czech taggers follow the traditional model, which (in our opinion, artificially) discounts most PoS ambiguities.

MSD is the biggest problem. Mea culpa, in the paper I didn't properly distinguish between tagging and MSD.

Disambiguation

Rule Based Machine Translation: It doesn't work, but we know why.

Disambiguation: Cascades of failure

Poor disambiguation is the root of most errors:

- Direct source of errors
- Leads to chunking errors
- Leads to chunk agreement errors

What's Missing?

- Fine grained disambiguation
- Look ahead in chunking (partly done)
- Multiword units with multiply inflected items

Deliberately Abrupt Ending

I really have to shut up... I could talk about this forever. Questions?

One Last Thing...

Joanna could not join us, as she recently started a new job. Her work on this project contributed to her successful job application; I hope you'll join me in congratulating her.